Manuel Soares Lourenço

Extracção de palavras compostas por bootstrapping

Universidade da Beira Interior

Departamento de Informática

Julho 2009

Manuel Soares Lourenço

Extracção de palavras compostas por bootstrapping

Tese submetida ao Departamento de Informática para o preenchimento dos requisitos para a concessão do grau de Mestre efectuada sob a supervisão do Doutor Gaël Harry Dias, Professor Assistente no Departamento de Informática da Universidade da Beira Interior, Covilhã, Portugal

Universidade da Beira Interior
Departamento de Informática
Julho 2009

Resumo

Nesta dissertação foi proposto um novo método, que altera o funcionamento de um sistema existente para extracção de palavras compostas. Este sistema, o SENTA tem uma falha, e este novo método tem por objectivo a correcção dessa falha, extraindo assim palavras compostas que não seriam extraídas pelo SENTA normal. Usando um algoritmo de bootstrapping para fazer o sistema SENTA trabalhar de forma recursiva, alterando o corpus a cada iteração.

Abstract

In this master thesis was proposed a new method, to change the way, an existing system to extract multiword units works. This system, named SENTA has a fail, and this method was built with the objective to correct it, extracting new multiword units that the normal SENTA don't extract. Making use of a bootstrapping algorithm to make the system SENTA working recursively changing the corpus used in each iteration.

Agradecimentos

Os primeiros agradecimentos vão para os meus pais por serem eles os principais financiadores e por me permitirem chegar até esta etapa, contando sempre com o apoio de ambos para as minhas decisões pessoais.

Ao Doutor Gaël Harry Dias por me ter apresentado esta proposta de tese e pela orientação prestada, os meus agradecimentos.

Finalmente a todos os meus amigos e colegas de curso, que também contribuíram sempre quando necessitei deles, quer em momentos de trabalho, quer em momentos de lazer, o meu obrigado a todos.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A informação digital tem sofrido um enorme crescimento nas últimas décadas, o que tem promovido que o desenvolvimento de uma área de pesquisa de informação de documentos, a IR[1] seja extremamente necessário, novas abordagens e novos algoritmos têm ajudado a melhorar a forma como se classifica, filtra, traduz e se resume documentos, tornando esta área cada vez mais complexa, mas também mais eficiente e mais automatizada. Ferramentas de extracção automática de informação de documentos, como é exemplo a extracção automática de palavras compostas surgem para responder a estas necessidades.

Nesta dissertação abordar-se-á um novo método de extracção de palavras compostas que combina um algoritmo já existente para este fim, actualmente conhecido como SENTA[2] [4] e um método de programação, o *bootstrapping*, que consiste essencialmente num método recursivo que vai alterando o *corpus*[3] a cada iteração. Um dos principais problemas do SENTA é uma restrição imposta pelo algoritmo *GenLocalMax* que é um algoritmo de aquisição que explicar-se-á detalhadamente no capítulo 3, esta restrição impossibilita a extracção de uma palavra composta de tamanho *n* se extrair uma correspondente (que contenha as mesmas palavras) de tamanho *n-1*. A introdução do *bootstrapping* procura resolver este problema do SENTA para obtenção de palavras compostas que não seriam extraídas utilizando o SENTA normalmente.

O capítulo 2, Extracção de palavras compostas, explica em que contexto o nosso problema se enquadra na área da IR, analisando a evolução da extracção de palavras compostas, enunciando trabalhos relacionados.

[1] do inglês *Information Retrieval*

[2] Software for the Extraction of N-ary Textual Associations

[3] Conjunto de textos utilizados para análise, retirados de um conjunto disponibilizado pela Reuters em 2000

O capítulo 3, Senta, explica o funcionamento do software SENTA desenvolvido pelo Alexandre Gil como complemento da sua dissertação: "Extracção eficiente de padrões textuais utilizando algoritmos e estruturas de dados avançadas" [7], que foi baseada no trabalho teórico de Gaël Dias, Sylvie Guilloré e José Gabriel Lopes em "Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora" [4].

O capítulo 4, Senta por bootstrapping, explica o funcionamento do método proposto por esta dissertação para combinar o SENTA com o *bootstrapping*.

O capítulo 5, Resultado, mostra os resultados obtidos e as comparações entre os dois métodos testados e nas várias configurações que permitem.

Finalmente, o último capítulo, mostra as conclusões chegadas nesta nova abordagem para a extracção de palavras compostas.

Capítulo 2

Extracção de palavras compostas

A expansão da *world wide web* e um uso cada vez mais comum de documentos em formato digital em detrimento de documentos físicos, como livros, revistas ou jornais tem provocado um crescimento exponencial de informação em formato digital, actualmente existe uma quantidade imensurável de documentos que necessitam de ser analisados, classificados, filtrados, etc. tudo automaticamente, pois devido a esta imensa quantidade tornou-se humanamente impossível. Esta realidade levou ao crescimento de métodos automáticos de selecção, tratamento e classificação de documentos, que cada vez mais necessitam de algoritmos rápidos e eficientes, e ao desenvolvimento de uma área de pesquisa de informação em Documentos, a *IR*, uma área que no passado estava mais centrada na indexação e na procura de documentos úteis numa colecção e que hoje inclui uma pesquisa em modelação, classificação e filtragem de documentos, interfaces com o utilizador, visualização de dados, etc [1]. Os grandes responsáveis pela evolução desta área de pesquisa de informação em documentos, os motores de busca, o *Google*, o *Yahoo*, e muito recentemente o *bing* da *Microsoft*, são os principais responsáveis pela investigação acelerada nesta área, em grande parte devido à concorrência que existe entre eles.

Uma das áreas ligadas a esta evolução do tratamento automático de documentos é a extracção automática de palavras compostas, palavras compostas são grupos de palavras que ocorrem frequentemente juntas e que têm um significado diferente do que ocorressem separadas, bons exemplos deste tipo de grupos de palavras serão por exemplo, nomes compostos (*Presidente da República*), verbos compostos (*Tenho feito*), locuções adverbiais(*de modo algum*), locuções preposicionais (*a respeito de*) ou locuções conjunctivas (*a fim de que*). Esta área não está muito desenvolvida e são ainda poucos os métodos com resultados que possam ser considerados muito bons. Segundo a comunidade cientifica

[5] existem três abordagens de extrair palavras compostas, uma primeira usando técnicas baseadas em métodos linguísticos, por exemplo etiquetagem morfossintática [1] ou utilização de padrões ou modelos linguísticos, uma segunda usando métodos puramente estatísticos onde a extracção das palavras compostas é um processo totalmente independente da lingua dos documentos, e uma terceira abordagem usando um misto das duas abordagens anteriores, onde se procura encontrar certos padrões textuais. A primeira e a terceira abordagens são dependentes da língua e obrigam à existência de bases de dados, actualizadas, de padrões linguísticos [6], [8] e ao desenvolvimento de medidas e dos respectivos limiares de aceitação.

Um bom exemplo para um método com bons resultados e o qual analisaremos nesta dissertação, é o *SENTA* [4], um método baseado em estatísticas para uma extracção em massa de palavras compostas, que usa uma medida de associação que mede a força entre as palavras de cada grupo, a *Expectativa Mútua*, e um algoritmo de selecção, o *GenLocalMax*. Este método tem uma boa taxa de eficiência para a extracção de palavras compostas, compostas por duas palavras, um dos pontos chaves para o estudo realizado nesta dissertação reside nesta eficiência, o objectivo compreende-se em utilizar o Senta por *bootstrapping* ou por outras palavras, utiliza-lo uma forma recursiva, utilizando-o várias vezes para um *corpus* que será substancialmente diferente em cada iteração dessa recursividade, de uma forma mais prática o que se propõe é utilizar o *SENTA* para extrair palavras compostas e substituir essas palavras compostas num novo corpus para posterior reutilização por uma palavra só, que representará esse grupo, ou seja a cada iteração o *SENTA* interpretará essa palavras compostas como sendo uma só palavra. Exemplificando: utilizando o *SENTA* para retirar palavras de tamanho 2 retiramos a seguinte palavra composta [A B] do seguinte corpus [A B C A D A B C A B] numa primeira iteração, modificando o corpus para uma segunda iteração teremos o seguinte corpus [A_B C A D A_B C A_B] que será utilizado novamente para o *SENTA* que detectará a palavra composta [A_B C] como sendo de tamanho 2, mas que na realidade será de tamanho 3. Outro ponto chave para este estudo ao contrário do primeiro utilizaremos uma falha e não uma eficiência do *SENTA*, uma falha que essencialmente advém do algoritmo de selecção, o *GenLocalMax*, esta falha é relativa ao facto de este algoritmo quando detecta um máximo local para uma palavra composta de tamanho *n* descarta a possibilidade de extrair as suas palavras compostas vizinhas, ou seja, as palavras compostas com tamanhos *n-1* e *n+1* que contenham as mesmas palavras da que tem tamanho *n*.

[1] Classificação das palavras segundo a sua morfologia sintáctica

Capítulo 3

Senta

O Senta é um software criado para a extracção automática de palavras compostas que usa um sistema baseado em estatísticas e que é independente da linguagem do corpus. A versão do Senta utilizada nessa dissertação é a versão desenvolvida por Alexandre Gil para a sua dissertação de Mestrado, "Extracção Eficiente de padrões textuais utilizando algoritmos e estruturas de dados avançadas". [7] O Senta consiste num sistema que constrói listas com n-gramas e a frequência em que cada um ocorre no corpus, atribuindo um valor baseado nessa frequência definindo uma medida de associação entre as palavras do n-grama, a *Expectativa Mútua*, e selecciona desse grupo as palavras compostas utilizando um processo de aquisição, o *GenLocalMax* que utiliza os máximos locais para seleccionar os n-gramas.

3.1 N-gramas

Os n-gramas são grupos de palavras que respeitam a ordem e a posição pelas quais estas ocorrem no corpus, estas podem ser contíguas ou não-contíguas, são gerados fazendo uma leitura ao corpus por grupos de M palavras, e gerados de cada um desses grupos, os sub-gramas possíveis, por exemplo se o nosso corpus for [A B C D E F], então [A B C], [C _ D], [_ D E _ F] serão exemplos de n-gramas que poderão ser criados. Os n-gramas serão representados nos próximos capítulos com a forma p11w1,p12w2,p13w3, ... , p1iwi onde p1i representa a distância entre a palavra i e a palavra 1, nesta representação foi escolhido o primeiro elemento como o elemento pivot, mas qualquer elemento poderia ser escolhido para tal.

3.2 Expectativa Mútua

A expectativa mútua é uma medida de associação para n-gramas, que permite identificar a força de ligação entre as palavras de um n-grama, avaliando o impacto da perda de cada uma das palavras no valor do conjunto. Outras medidas de associação utilizadas até então são insatisfactórias, pois elas só avaliam o grau de associação entre duas palavras enquanto a expectativa mútua avalia para N palavras.

Segundo [3], a expectativa mútua é calculada da seguinte forma:

$$ME(w) = p(w) \times NE(w)$$

onde *(*w) é um *n-grama*, p(w) é a probabilidade de ocorrência do *n-grama w* no *corpus* em análise e NE(w) é a medida normalizada da expectativa associada a *w*, ME vem do inglês *Mutual Expectation* e NE *Normalized Expectation*.

A NE(w) associada a um *n-grama* é definida como sendo a expectativa média da ocorrência duma palavra numa determinada posição, conhecendo-se a ocorrência das outras palavras desse *n-grama*, igualmente restringidos às suas posições [3], e é calculada da seguinte forma:

$$NE(w) = \frac{p(w)}{FPE(w)}$$

onde FPE(w) do inglês *Fair Point of Expectation* é a média das probabilidades de ocorrência de cada sub-gram de *w* no *corpus* e é calculado da seguinte forma:

$$FPE(w) = \frac{p([w_2 p_{23} w_3 p_{24} w_4 ... p_{2t} w_t]) + \sum_{i=2}^{t} p([w_1 p_{12} w_2 ... \hat{p_{1i}} \hat{w_i} ... p_{1t} w_t])}{t}$$

onde o símbolo "$\hat{}$" representa a palavra a eliminar em cada passo, para a formação de cada sub-grama. O valor de *t* corresponde ao número de *sub-gramas* válidos gerados a partir do *n-grama* w.

3.3 Algoritmo GenLocalMax

O *GenLocalMax* é um algoritmo de selecção puramente estatístico independente da linguagem do *corpus*. O *GenLocalMax* necessita de uma medida de associação crescente, ou

seja, onde os valores mais altos correspondem a casos mais relevantes, neste caso a medida de associação que é utiliza é a atrás referida, a Expectativa Mútua.

Para que um *n-grama* seja considerado uma palavra composta, este terá de ser um máximo local, ou seja sendo *w* um *n-grama* terá de verificar as seguintes condições [2]:

$$\forall x \varepsilon \Omega_{n-1} \; \forall y \varepsilon \Omega_{n+1}$$

$$tamanho(w) = 2 \; e \; g(w) > g(y) \; ou$$

$$tamanho(w) > 2 \; e \; g(x) \leq g(w) \; e \; g(w) > g(y)$$

onde tamanho() representa o número de palavras de cada *n-grama*, g() a função da medida de associação, *w* o n-grama, *x* um *(n-1)-grama* e *y* um *(n+1)-grama*.

Capítulo 4

Senta por Bootstrapping

Explicado o funcionamento do *Senta* no capítulo anterior, pode-se passar a este novo capítulo que demonstra o trabalho prático realizado, abordando um novo sistema que combina o Senta com um método estatístico, o *bootstrapping*, um método que vai alterando o *corpus* e vai executando várias vezes o *Senta* com esse novo *corpus*, ou seja, o *corpus* a ser utilizado vai sendo refinado e reutilizado em cada nova iteração, este processo é repetido até o Senta não extrair novas palavras compostas do *corpus* em análise.

O *Senta* utilizado está modificado para assimilar os *underscores* "_" como sendo letras normais, assim ele considera por exemplo seguinte palavra composta: "Presidente_da_república" como sendo só uma única palavra, e é esta a chave para a construção dos novos *corpus*.

O *corpus* utilizado para os testes é um *corpus* público na língua inglesa lançado pela *Reuters* em 2000, que é um conjunto de ficheiros de texto de um apanhado de inúmeras notícias. Nesta dissertação é utilizada uma pequena parcela desse *corpus* contendo aproximadamente 1.2 milhões de palavras.

4.1 Arquitectura do sistema

O sistema como se pode ver na figura é constituído por 4 partes distintas, o *SENTA* que é a parte mais importante do sistema onde é feita a selecção das palavras compostas, o *MwuExtract* que faz uma selecção das palavras compostas retiradas do resultados do *SENTA*, o *CorpusBuilder* que cria um novo *corpus*, e uma parte de verificação da existência de palavras compostas, que fará o programa parar quando não encontrar mais.

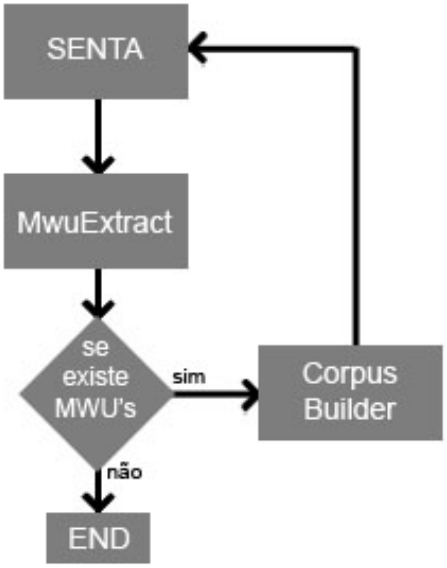

Figura 4.1: Funcionamento do Senta por *bootstrapping*

4.2 MwuExtractor

o *MwuExtractor* selecciona as palavras compostas que tenham um tamanho igual ou inferior ao pretendido, e ignora todas as palavras compostas não contíguas, pois só se testará este sistema para palavras compostas contíguas.

4.3 CorpusBuilder

O *CorpusBuilder* utiliza a lista criada pelo *MwuExtractor* e analisa a existência de cada uma das palavras compostas seleccionadas em todos os ficheiros do *corpus* substituindo cada palavra composta encontrada por uma equivalente, mas separada por um *underscore* na vez de um espaço, tendo em conta que o *SENTA* está preparado para ignorar os *underscores*, que os considera como se fossem uma letra, fazendo com que as palavras compostas que foram modificadas no *corpus* sejam interpretadas na próxima iteração do sistema como se fossem uma única palavra.

Capítulo 5

Resultados

Neste capítulo apresenta-se e analisa-se os resultados dos testes realizados com o Senta por bootstrapping e com o Senta Normal. O objectivo é comparar os resultados obtidos em ambos os casos, para saber ate que ponto a recursividade do Senta por bootstrapping consegue resolver a falha principal do Senta Normal já referida em capítulos anteriores, um falha que advém do algoritmo *GenLocalMax* que não permite a extracção de uma palavra composta com *n* palavras se já tiver sido extraída uma palavra compostas com *n-1* dessas palavras.

Como foi explicado no capítulo anterior o Senta por *bootstrapping* utiliza duas variáveis que podem ser ajustadas para obter resultado diferentes, estas duas variáveis que serão enunciadas por *N* e *M*, e são respectivamente, o tamanho dos n-gramas utilizados pelo Senta e o tamanho máximo das palavras compostas que no novo corpus construído serão consideradas como uma única palavra. Na tabela 5.1 é demonstrado como são formados os n-gramas, seleccionando uma palavra e as suas vizinhas, por essa razão é que são utilizados os valores ímpares 3, 5 e 7 para o *N*, porque as palavras vizinhas seleccionadas são sempre na mesma quantidade para a esquerda e direita.

Tabela 5.1: N-gramas utilizados pelo Senta

N = 3	-1 palavra +1
N = 5	-2 palavra +2
N = 7	-3 palavra +3

5.1 Número de palavras extraído

Nesta primeira secção, são comparados os valores obtidos de três testes diferentes do Senta por bootstrapping e três testes do Senta Normal, os testes escolhidos para esta comparação inicial são para o Senta por bootstrapping com valores de *M* e *N*, de 2/3, de 4/5 e de 6/7, onde o primeiro valor é o *M* e o segundo o *N* respectivamente. O valor de *M* é sempre uma unidade inferior a *N*, porque para n-gramas de tamanho *I*, as palavras compostas que o Senta extrai são no máximo de tamanho *I-1*, logo, os resultados para valores de *M* iguais ou superiores a *N* serão iguais aos resultados para os exemplos mostrados.

Nas seguintes tabelas 5.2 e 5.3 estão as quantidades de palavras compostas extraídas, utilizando um *corpus* de 1,2 milhões de palavras, os valores entre os parêntesis são as quantidades de palavras compostas extraídas a cada iteração do sistema, pode-se verificar que a quantidade de palavras compostas extraídas na primeira iteração de cada um dos testes do Senta por *bootstrapping* são as mesmas que no Senta Normal para os mesmos valores de *N*, foi verificado que são exactamente as mesmas palavras extraídas.

Tabela 5.2: Resultados do Senta por *bootstrapping*

M	N	Quantidade de Palavras compostas extraídas
2	3	168 (139 + 25 + 4)
4	5	385 (319 + 51 + 14 + 1)
6	7	390 (327 + 52 + 9 + 1 + 1)

Tabela 5.3: Resultados do Senta Normal

N	Quantidade de Palavras compostas extraídas
3	139
5	319
7	327

5.2 Várias formas de aplicar a recursividade

Na secção anterior comparamos as diferenças entre os resultados obtidos no Senta Normal e no Senta por bootstrapping para os mesmos valores de *N* e para um *M* com uma unidade apenas, inferior em relação ao *N*, nesta secção vamos comparar os resultados obtidos para outros valores do *M* com uma maior diferença em relação ao *N*.

Podemos verificar na tabela 5.4 várias formas de aplicar a recursividade ao Senta por *bootstrapping*, quanto maior for a diferença entre o *M* e o *N* menor é quantidade me palavras compostas extraídas, logo os resultados para essas diferenças maiores do que uma unidade, podem ser desprezadas, pois o objectivo é encontrar palavras compostas que á partida o Senta ignore, e não ignorar as que ele encontre.

Tabela 5.4: Resultados do Senta por *bootstrapping* para vários valores de *M*

M	N	Quantidade de Palavras compostas extraídas
2	3	168 (139 + 25 + 4)
2	5	130 (116 + 14)
3	5	333 (258 + 57 + 16 + 2)
4	5	385 (319 + 51 + 14 + 1)
2	7	122 (107 + 15)
3	7	319 (245 + 57 + 15 + 2)
4	7	365 (299 + 56 + 10)
5	7	379 (317 + 52 + 9 + 1)
6	7	390 (327 + 52 + 9 + 1 + 1)

5.3 Caracterizar o Senta por Bootstrapping em relação ao Senta Normal

Como já foi explicado anteriormente o ponto fraco do Senta é falhar na extracção de uma palavra composta de tamanho *I* quando extrai uma palavra composta de tamanho *I-1* com palavras comuns. Por exemplo se o Senta extrair a palavra composta “Carbon dioxide” já não consegue extrair a palavra “Saturated carbon dioxide”. Por esta razão é mostrado na tabela 5.5 a quantidade de palavras que o Senta por *bootstrapping* extraiu que não

seriam extraídas pelo SENTA normal. Mais uma vez é verificado que a melhor solução de configuração para as variáveis de *M* e *N* é usar a diferença de uma unidade ou seja *M* = *N* - 1.

Tabela 5.5: Quantidade de palavras compostas com tamanho *I* extraídas pelo Senta por *bootstrapping* em que a palavra composta com tamanho *I-1* foi extraída pela Senta Normal

M	N	Quantidade de Palavras compostas extraídas
2	3	8
2	5	6
3	5	12
4	5	14
2	7	6
3	7	16
4	7	15
5	7	15
6	7	18

5.4 Precisão da Extracção

A precisão de extracção, ou seja, a quantidade de palavras compostas extraídas que realmente são consideradas como palavras compostas, que têm um significado diferente juntas do que teriam separadas, é mostrada nas seguintes tabelas, na tabela 5.6 onde são mostradas as frequências de acertos para os vários casos, e na tabela 5.7 onde são mostradas as percentagens. Só mostramos a precisão de extracção para os casos em que o *M* é igual a *N* - 1, pois como verificamos nas duas secções anterior, os resultados não acrescentam nada de novo, também só mostramos os resultados para o Senta por *bootstrapping* pois como também já foi mostrado anteriormente os valores do SENTA Normal são os mesmos da primeira iteração do Senta por *bootstrapping*.

Tabela 5.6: Frequência de precisão de extracção do Senta por *bootstrapping*

M	N	Frequência da precisão
2	3	86 (78 + 7 + 1)
4	5	136 (123 + 8 + 5)
6	7	129 (111 + 15 + 3)

Tabela 5.7: Percentagem de precisão de extracção do Senta por *bootstrapping*

M	N	Percentagem de precisão
2	3	51,1 %
4	5	35,3 %
6	7	33,1 %

5.5 Análise qualitativa

Nesta secção estão alguns exemplos de palavras compostas extraídas pelo a partir da segunda iteração, palavras compostas que não seriam extraídas pelo SENTA Normal, os exemplos de teste nesta tabela 5.8 continuam a ser os mesmos três que foram abordados nas secções anteriores para os quais verificou-se que os resultados seriam melhores.

Tabela 5.8: Palavras Compostas extraídas pelo Senta por *bootstrapping*

M	N	Palavras compostas extraídas
2	3	“ahead of the”; “Financial Times”; “should be”.
4	5	“by the end of the”; “in the first half of”; “Olympic Games”.
6	7	“New York Mercantile Exchange”; “have been”; “Agriculture Department”.

Capítulo 6

Conclusão

Este método para extrair palavras compostas apresentado é um método pesado e que necessita de bastante tempo de processamento comparado com o SENTA normal, este poderá ser optimizado visto que não era objectivo desta dissertação implementar um software optimizado e sim funcional que esclarecesse se realmente conseguiria responder aos objectivos de corrigir a falha, que foi explorada, do SENTA normal. Depois de analisados os resultados obtidos por este método, podemos concluir que o SENTA por bootstrapping consegue resolver a falha detectada no SENTA e extrair palavras compostas que não seriam extraídas, embora sejam em pequenas quantidades poderão ser suficientes para fazer do SENTA por bootstrapping um sistema mais completo, não deixando possíveis palavras compostas por extrair.

Bibliografia

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, May 1999.

[2] Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *EPIA*, pages 113–132, 1999.

[3] Gaël Dias, Sylvie Guilloré, Jean claude Bassano, José Gabriel, and José Gabriel Pereira Lopes. Combining linguistics with statistics for multiword term extraction: A fruitful association? 2000.

[4] Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. 1999.

[5] Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. Multilingual aspects of multiword lexical units. 1999.

[6] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured texts, 2000.

[7] Alexandre Nuno Capinha Gil. Extracção eficiente de padrões textuais utilizando algoritmos e estruturas de dados avançadas. Master's thesis, Universidade Nova de Lisboa, 2002.

[8] Satoru Ikehara, Satoshi Shirai, and Hajime Uchino. A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In *Proceedings of the 16th conference on Computational linguistics*, pages 574–579, Morristown, NJ, USA, 1996. Association for Computational Linguistics.